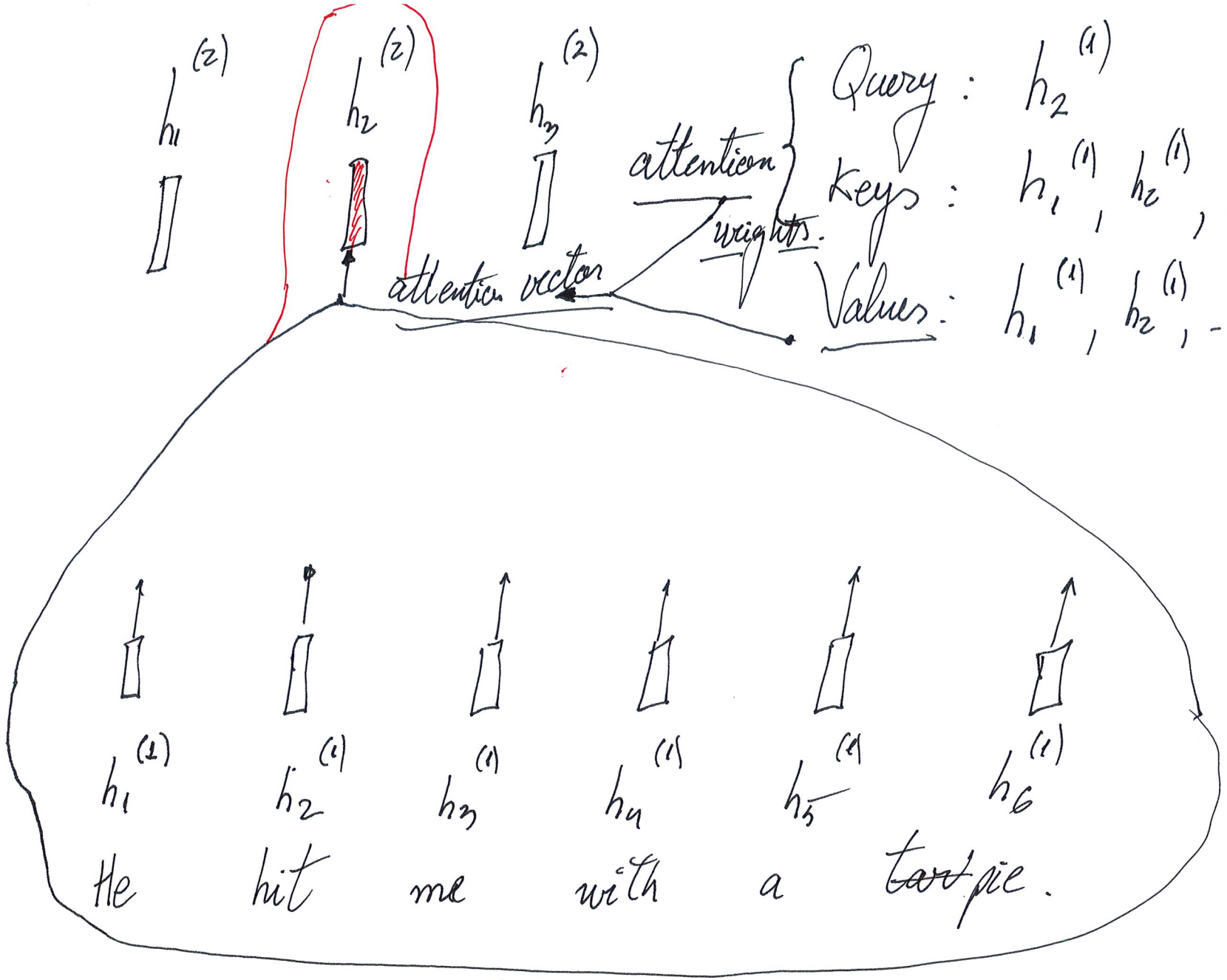

$$\begin{cases} W_Q h_2^{(1)} \longrightarrow \text{query vector} \\ W_K h_1^{(1)}, W_K h_2^{(1)}, \dots, W_K h_6^{(1)} \longrightarrow \text{key vector.} \end{cases}$$

Attention weights :

$$\left[W_Q h_2^{(1)}\right]^T \left(W_K h_1^{(1)}\right) \rightarrow \tilde{a}_1 / \sqrt{d_k}$$

$$\left[- , - \right]^T \left(W_K h_2^{(1)}\right) \rightarrow \tilde{a}_2 / \sqrt{d_k}$$

$$\vdots$$

$$\left[- 6 - \right]^T \left(W_K h_6^{(1)}\right) \rightarrow \tilde{a}_6 / \sqrt{d_k}$$

$$\left\{ \tilde{a}_1, \tilde{a}_2, \ldots, \tilde{a}_6 \right\} \xrightarrow{\text{softmax}} \left\{ a_1, a_2, \ldots, a_6 \right\}$$

$$= a_1 \cdot \left(W_V h_1^{(1)}\right) + a_2 \left(W_V h_2^{(1)}\right) + \ldots + a_6 \left(W_V h_6^{(1)}\right)$$

attention vector for one attention head

(self)